Not for distribution

Notes for Plant Breeding 150 (Algebra)

BU-49-M J. E. Dowd Fall, 1953

Summation

Suppose X is a variable which is observed in a population. Suppose as a result of making n observations on the variable X we arrive at a series of individual observations or variates $x_1, x_2, x_3, \ldots, x_i, \ldots, x_n$. For example X might be the yield in bushels of a plot of wheat. Then $x_1, x_2, \ldots, x_i, \ldots, x_n$ would be the yields in bushels from the 1st, 2nd, i-th, n-th plots respectively. The subscripts 1, 2, ..., n are used merely to identify the various plot yields uniquely, and there is no ordering implied, i.e., x_1 is not necessarily larger or smaller than x_2 etc. x_i is the result of an observation on the i-th individual or object on the population, and the notation x_i or (x_i) i=1, ... n, is often used to denote the series of n observations made.

One operation often performed in statistics is that of summation, denoted generally by Σ or S. $\sum_{i=1}^{n} x_i$ means "Add up all quantities like x_i which are formed by giving i the values of every positive integer from i=1 to i=n inclusive." Thus $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_i + \ldots + x_n$. It is important to note that the index i must take on all integral values between its lower limit given beneath the summation sign and its upper limit given above the summation sign.

Examples:

1. $\sum_{i=1}^{n} x_i y_i = x_1y_1 + x_2y_2 + \ldots + x_iy_i + \ldots + x_ny_n$

2. $\sum_{j=1}^{n} (-1)^j p_{j-1} = -p_0 + p_1 - p_2 + \ldots + (-1)^i p_{i-1} + (-1)^n p_{n-1}$

3. $\sum_{i=1}^{4} (y_i + 5) = (y_i + 5) + (y_2 + 5) + (y_3 + 5) + (y_4 + 5) = y_1 + y_2 + y_3 + y_4 + 20$

4. $\sum_{j=1}^{n/2} y_{2j-1} = y_1 + y_3 + y_5 + \ldots + y_{n-1}$ n is an even integer.

Note that while i must take on all integral values between

its upper and lower limits that x_i does not necessarily have to take on all its calues in the summation. Example 4 above is a case where, while i takes on all integral values between i and n/2, x_i takes on those values between 1 and n which are characterized by an odd subscript.

Note also that if the variable x is to take on the particular values 1, 2, 3, 4, ... n instead of the general values x_1, x_2, x_3, x_n we may write $\sum_{x=1}^{n} x = 1 + 2 + 3 + \ldots + n$

$$\sum_{x=1}^{n} x^2 = 1^2 + 2^2 + 3^2 + \ldots + n^2$$

We now state 3 useful theorems which the student may easily verify by expanding the expressions and regrouping the terms.

Theorem I: $\sum_{i=1}^{n}(x_i + y_i - z_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} z_i$

Theorem II: $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$ where c is a constant.

Theorem III: $\sum_{i=1}^{n} c = nc$ where c is a constant.

Example: $\sum_{i=1}^{n}(x_i - c) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} c$ by Theorem I

$= \sum_{i=1}^{n} x_i - nc$ by Theorem III.

If we define the arithmetic mean as

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_i + \ldots x_n}{n}$$

then we may write $\bar{x}$ in our shorthand notation as

$$\bar{x} = 1/n \sum_{i=1}^{n} x_i$$

It is easy to see that the sum of the deviates from the mean i iiis equal to zero. We denote an individual deviation from the

mean as $(x_i - \bar{x})$.

Then $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}$

$= \sum_{i=1}^{n}x_i - n\bar{x}$ since $\bar{x}$ is a constant

but $\bar{x} = 1/n \sum_{i=1}^{n}x_i$ or $n\bar{x} = \sum_{i=1}^{n}x_i$

therefore $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}x_i = 0.$

If we have n observations $x_1, x_2, \ldots x_n$, such that f_1 of the observations have a common value, say y_1, f_2 observations have a common value, say y_2, f_k observations have a common value y_k, then we can write

$$\bar{x} = 1/n \sum_{i=1}^{n}x_i = \frac{\sum_{i=1}^{k}f_i y_i}{\sum_{i=1}^{k}f_i} \text{ or } 1/n \sum_{i=1}^{k}f_i y_i \text{ since } \sum_{i=1}^{k}f_i = n.$$

For example, $n = 8$, $x_1 = 7$, $x_2 = 7$, $x_3 = 4$, $x_4 = 7$, $x_5 = 5$, $x_6 = 4$, $x_7 = 6$, $x_8 = 5$.

$y_1 = 7$	$f_1 = 3$	$y_1f_1 = 21$
$y_2 = 4$	$f_2 = 2$	$y_2f_2 = 8$
$y_3 = 5$	$f_3 = 2$	$y_3f_3 = 10$
$y_4 = 6$	$f_4 = 1$	$y_4f_4 = 6$
	8	45

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots . x_8}{8} = \frac{7 + 7 + \ldots + 5}{8} = \frac{45}{8} = 5\ 5/8$$

$$= \frac{y_1f_1 + y_2f_2 + \ldots + y_4f_4}{f_1 + f_2 + \ldots + f_4} = \frac{45}{8} = 5\ 5/8$$

Coding: It is often convenient in practice to deal with not the original data but with values derived from the original

data by coding the value of each operation. We may code by subtracting a constant from each operation:

$$x_i' = x_i - x_o$$

where x_i' is the new coded value,
x_i is the original observation
x_o is a constant.

$$\sum_{i=1}^{n} x_i' = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_o = \sum_{i=1}^{n} x_i - nx_o$$

or

$$\bar{x} = 1/n \sum_{i=1}^{n} x_i' = 1/n\left(\sum_{i=1}^{n} x_i - nx_o\right) = \bar{x} - x_o$$

or $\bar{x} = \bar{x}' + x_o$

It can be shown also that if we code by

$$\bar{x} = \frac{x_i - x_o}{c} \qquad \text{where } c \neq 0$$

then $\bar{x} = c\bar{x}'' + x_o$

Problem To show that $\sum_{i=1}^{n} x = 1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$

Solution: Consider $x^2 - (x-1)^2 = x^2 - x^2 + 2x - 1 = 2x - 1$

Sum both sides from 1 to n.

$$\sum_{i=1}^{n} \left[x^2 - (x-1)^2\right] = \sum_{i=1}^{n} (2x - 1) = 2\sum_{i=1}^{n} x_i - n$$

Performing the indicated sum on the left member we have:

$$\begin{aligned} &1^2 - 0^2 \\ &+ 2^2 - 1^2 \\ &+ 3^2 - 2^2 \\ &\cdot \\ &\cdot \\ &\cdot \\ &+ n^2 - (n-1)^2 \end{aligned}$$

$$n^2 = 2\sum_{i=1}^{n} x_i - n$$

or transposing and dividing

$$\sum_{i=1}^{n} x_i = \frac{n^2 + n}{2} = \frac{n(n+1)}{2}$$

Exercises

1. Write in expanded form:

(a) $\sum_{i=1}^{n} x_i^2$

(b) $\sum_{i=n_1+1}^{n_1+n_2} x_i$

(c) $\sum_{i=1}^{4} (x_i - \bar{x})^2$

(d) $\sum_{j=1}^{5} (-1)^{j-1} jx^{2j}$

2. Express $a_o + a_1x + a_2x^2 + a_3x^3 + \ldots + a_nx^n$ in Σ notation.

3. Prove: (a) $\sum_{i=1}^{N} (x_i + 1)^2 = \sum_{i=1}^{N} x_i^2 + 2\sum_{i=1}^{N} x_i + N$

(b) $\sum_{x=0}^{N} x(x - 1)p = \sum_{x=2}^{N} x(x - 1)p$

4. Distinguish between $\sum_{i=1}^{n} x_i y_i$ and $\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$.
Write each in expanded form.

5. Using the identity $x^3 - (x - 1)^3 = 3x^2 - 3x + 1$,

show that $\sum_{x=1}^{n} x^2 = \frac{m(n + 1)(2n + 1)}{6}$

6. If the mean of a set of n_1 variates is $\bar{x}_1$ and the mean of another set of variates of the same variable is $\bar{x}_2$, show that the mean $\bar{x}$ of the combined sets is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{N} \qquad \text{where } N = n_1 + n_2.$$

Generalize this result to k sets of variates of the same variable x, where $\bar{x}_1 \ldots \bar{x}_k$ are the means of k sets and $n_1 \ldots n_k$ are the size of the sets and $n_1 + n_2 + \ldots + n_k = N$. Express your answer in Σ notation. What does the formula reduce to if all sets are of the same size? i.e., $n_i = n$ $\quad i=1 \ldots k$

The variance of a sample of size n is defined as

$$s^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right]$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

1. To derive the computational formula:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i - \frac{(\Sigma x_i)^2}{n}\right] = \frac{1}{n-1}\left[\Sigma x_i^2 - n(\bar{x})^2\right]$$

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})$$

$$= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2$$

$$= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2 \qquad \text{(Using Theorems I, II, and III)}$$

but $\bar{x} = \frac{\sum_{i=1}^{n}x}{n}$ or $n\bar{x} = \sum_{i=1}^{n}x_i$

$$\therefore (n-1)s^2 = \sum_{i=1}^{n}x_i^2 - \frac{2\sum_{i=1}^{n}x_i}{n}\sum_{i=1}^{n}x_i + n\left(\frac{\Sigma x_i}{n}\right)^2$$

$$= \sum_{i=1}^{n}x_i^2 - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2 + \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2$$

$$= \sum_{i=1}^{n}x_i^2 - \left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2$$

$$\text{or } (n-1)s^2 = \sum_{i=1}^{n}x_i^2 - \frac{(n\bar{x})^2}{n}$$

$$= \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2$$

Problems (cont.)

7. Show that the mean of the first n integers is $\frac{n+1}{2}$ and the variance is $s^2 = \frac{1}{12} n(n+1)$.

8. If n_1, $\bar{x}_1$, and s_1^2 are the number, mean, and variance for one group of measures, and n_2, $\bar{x}_2$ and s_2^2 the number, mean, and variance for a second group, show that the variance of the group formed by combining the two groups is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n-2} + \frac{(n_1 - 1)(n_2 - 1)}{(n-2)^2} (\bar{x}_1 - \bar{x})^2$$

where $n = n_1 + n_2$

and s^2 = variance for the combined groups.

9. If X_i and Y_i i=1...n are two measurements and if $x_i = X_i - \bar{x}$ and $y_i = Y_i - \bar{y}$, show that

$$\sum_{i=1}^{n}(y_i - ax_i)^2 = (n-1)(a^2 s_x^2 + s_y^2) - 2a\sum_{i=1}^{n} x_i y_i$$

where a is a constant.

- -

Effect of coding on s^2:

Let $x_i'' = \frac{X_i - x_o}{c}$ or $X_i = cx_i'' + x_o$ and $\bar{x} = c\bar{x}'' + x_o$

then $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[cx_i'' + x_o - (c\bar{x}'' + x_o)\right]^2$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(cx_i'' + x_o - c\bar{x}'' - x_o)^2 = \frac{c^2}{n-1}\sum_{i=1}^{n}(x_i'' - \bar{x}'')^2$$

then if we define $s''^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i'' - \bar{x}'')^2$

we get

$$s^2 = c^2 s''^2$$

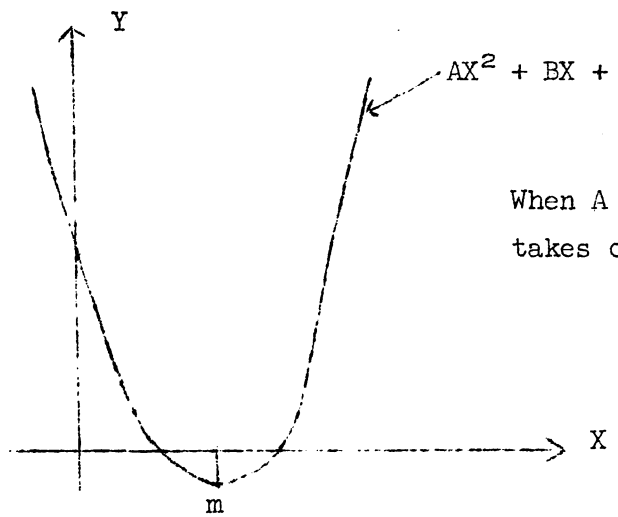
or $s = cs''$

Coding by adding or subtracting a constant value from each of the original observations has no effect upon the variance or standard deviation, but multiplication of the original observations will also multiply the standard deviation of the observations by the same amount.

To show that $\sum_{i=1}^{n}(x_i - a)^2$ takes on its minimum value when $a = \bar{x}$.

Method 1. Let $\sum_{i=1}^{n}(x-\bar{x})^2 = \sum_{i=1}^{n}(x-a+a-\bar{x})^2 = \sum_{i=1}^{n}(x-a)^2 + n(a-\bar{x})^2$ which is obviously at its smallest when $a = \bar{x}$ since the second positive term disappears. (see Chapter II, page 17.) A more general method is to show directly that the value of a which minimizes $\sum_{i=1}^{n}(x_i-a)^2$ is $a = \bar{x}$.

Consider first $(x_i-a)^2 = x_i^2 - 2ax_i + a^2$. This is of the general form $AX^2 + BX + C$ where $A = 1$, $B = -2a$, and $C = a^2$. If A is positive



When A is positive, parabola takes on lowest value at X = m.

$$AX^2 + BX + C = A(X + \frac{B}{2A})^2 + \frac{4AC - B^2}{4A}$$

since $A(X + \frac{B}{2A})^2 + \frac{4AC - B^2}{4A} = A(X^2 + \frac{B}{A}X^2 + \frac{B^2}{4A^2}) + \frac{4AC}{4A} - \frac{B^2}{4A}$

$= AX^2 + BX + \frac{B^2}{4A} + C - \frac{B^2}{4A} = AX^2 + BX + C.$

If A is positive, then $A(X + \frac{B}{2A})^2 + \frac{4AC - B^2}{4A}$ takes on its smallest value when $(X + \frac{B}{2A})^2 = 0$ or $(X + \frac{B}{2A}) = 0$ or $X = \frac{-B}{2A}$.

therefore $(x_i^2 - 2ax + a^2)$ takes on its smallest value when $x_i = \frac{2a}{2}$ or $x_i = a$.

$(x_1 - a)^2$ takes on its smallest value when $x_1 = a$
$(x_2 - a)^2$ " " " " " " " " $x_2 = a$
.
.
.
$(x_n - a)^2$ " " " " " " " " $x_n = a$

or adding $\sum_{i=1}^{n}(X_i - a)^2$ takes on its smallest value when $\sum_{i=1}^{n} X_i = na$

or $a = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{x}$.

Classification of a variable with respect to more than characteristic:

If we want to classify the variable X in a more explicit manner, we may use more than one subscript to denote the observations; e.g., the yield from the jth variety in the ith plot may be described as x_{ij}. If we were making observations on k different varieties and each variety was planted in n plots, then j = 1...k and i = 1...n. Altogether we would be making kn observations on X.

The values of the observations set out in a two way table with their sums and means would look as follows:

							Sum	Mean
x_{11}	x_{12}	x_{13}	...	x_{1j}	...	x_{1k}	$X_{1.}$	$\bar{x}_{1.}$
x_{21}	x_{22}	x_{23}	...	x_{2j}	...	x_{2k}	$X_{2.}$	$\bar{x}_{2.}$
x_{31}	x_{32}	x_{33}	...	x_{3j}	...	x_{3k}	$X_{3.}$	$\bar{x}_{3.}$
x_{i1}	x_{i2}	x_{i3}	...	x_{ij}	...	x_{ik}	$X_{i.}$	$\bar{x}_{i.}$
x_{n1}	x_{n2}	x_{n3}	...	x_{nj}	...	x_{nk}	$X_{n.}$	$\bar{x}_{n.}$
Sum $X.1$	$X.2$	$X.3$	...	$X.j$	...	$X.k$	$X..$	$\bar{x}..$
Mean $\bar{x}.1$	$\bar{x}.2$	$\bar{x}.3$	...	$\bar{x}.j$	...	$\bar{x}.k$	$\bar{x}..$	

Note that in this particular case the columns represent the varieties and the rows represent the plots.

The total for the j-th column is given by $X_{1j} + X_{2j} + \ldots + X_{nj}$ $= \sum_{i=1}^{n} X_{ij}$, and will be designated by X.j. We use a dot (.) in place of the subscript i to denote the fact that we have summed over the range of values taken on by i. The total for the i-th column would be $X_{i1} + X_{i2} + \ldots + X_{ik} = \sum_{j=1}^{k} X_{ij} = X_{i.}$ The total for the whole array is:

$$(x_{11} + x_{12} + \ldots + x_{1k}) + (x_{21} + x_{22} + \ldots + x_{2k}) + \ldots + (x_{n1} + \ldots + x_{nx})$$

$$= \sum_{j=1}^{k} X_{1j} + \sum_{j=1}^{k} X_{2j} + \ldots + \sum_{j=1}^{k} X_{ij} + \ldots + \sum_{j=1}^{k} X_{nj}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij}.$$ Also the total can be written as:

$$X.. = (x_{11} + x_{21} + x_{31} + \ldots + x_{n1}) + (x_{12} + x_{22} + \ldots + x_{n2})$$

$$+ \ldots + (x_{1j} + \ldots + x_{nj}) + (x_{1k} + \ldots + x_{nk})$$

$$= \sum_{i=1}^{n} X_{i1} + \sum_{i=1}^{n} X_{i2} + \ldots + \sum_{i=1}^{n} X_{ij} + \ldots + \sum_{i=1}^{n} X_{ik}$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n} X_{ij}.$$

It can readily be seen that similar relationships will hold for the means; e.g.

$$\bar{x}.. = \frac{1}{k} \sum_{j=1}^{k} \bar{x}.j = \frac{1}{n} \sum_{i=1}^{n} \bar{x}i. = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \bar{x}_{ij}$$

The variances within any row and any column are given by

$$s_i^2 = \frac{1}{k-1} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{i.})^2$$

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{.j})^2$$

Both of these estimate σ^2, the population variance.

The variances of the row and column means are:

$$s^2_{\bar{x}_{i.}} = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{x}_{i.} - \bar{x}..)^2$$

$$s^2_{\bar{x}_{.j}} = \frac{1}{k-1} \sum_{j=1}^{k} (\bar{x}_{.j} - \bar{x}..)^2$$

which estimate $\frac{\sigma^2}{k}$ and $\frac{\sigma^2}{n}$ respectively. (Note that each $\bar{x}_{i.}$ is a mean of k items and each $\bar{x}_{.j}$ is the mean of n items.)

Pooled variances within rows:

$$\frac{1}{k(n-1)} \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - \bar{x}_{i.})^2$$

Pooled variances within columns:

$$\frac{1}{k(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - x_{.j})^2$$

Total variance for all items:

$$\frac{1}{nk-1} \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}..)^2$$

This is also an estimate of σ^2.

Problem 10: Show that:

$$\frac{n_1 n_2}{n} (\bar{x}_1 - \bar{x}_2)^2 = n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2$$

where $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$ and $n_1 + n_2 = n$.

How is this simplified when $n_1 = n_2$?
If you then have k groups instead of 2 groups, generalize the formula for the pooled variances.

We will now show that the total sum of squares can be broken into two sums of squares, one the sum of squares between row means and the other the pooled sum of squares within rows. I.e.

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}..)^2 = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{i.})^2 + k \sum_{i=1}^{n} (\bar{x}_{i.} - \bar{x}..)^2$$

Write $X_{ij} - \bar{x}.. = x_{ij} - \bar{x}_{i.} + \bar{x}_{i.} - \bar{x}..$

$$\sum\sum_{ij}(X_{ij} - \bar{x}..)^2 = \sum\sum_{ij}[(x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}..)]^2$$

$$= \sum\sum_{ij}[(x_{ij} - \bar{x}_{i.})^2 + 2(x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}..) + (\bar{x}_{i.} - \bar{x}..)^2]$$

$$= \sum\sum_{ij}(x_{ij} - \bar{x}_{i.})^2 + \sum_{i}[(\bar{x}_{i.} - \bar{x}..)\sum_{j}(x_{ij} - \bar{x}_{i.})] + \sum\sum_{ij}(\bar{x}_{i.} - \bar{x}..)^2$$

since $(\bar{x}_{i.} - \bar{x}..)$ is a constant with respect to $\sum_j$. But note that

$$\sum_j(x_{ij} - \bar{x}_{i.}) = \sum_j x_{ij} - \sum_j \bar{x}_{i.} = k\bar{x}_{i.} - k\bar{x}_{i.} = 0 \quad \text{and}$$

$$\sum\sum_{ij}(\bar{x}_{i.} - \bar{x}..)^2 = \sum\sum_{ji}(\bar{x}_{i.} - \bar{x}..)^2 = k\sum_i(\bar{x}_{i.} - \bar{x}..)^2$$

$$\therefore \quad \sum\sum_{ij}(x_{ij} - \bar{x}..)^2 = \sum\sum_{ij}(x_{ij} - \bar{x}_{i.})^2 + k\sum_i(\bar{x}_{i.} - \bar{x}..)^2$$

Problem 11. Show that:

$$\sum\sum_{ij}(X_{ij} - \bar{x}..)^2 = \sum\sum_{ij}(X_{ij} - \bar{x}_{.j})^2 + n\sum_i(\bar{x}_{.j} - \bar{x}..)^2$$

Problem 12. Derive computational formulae for each sum of squares in 11.

Problem 13. What is the relationship between the degrees of freedom for the sums of squares in 11?

Expected values - Let X denote a variable which takes on a certain range of values or variates $X_1, X_2, \ldots X_N$. N may be a finite number if the population characterized by X is finite in size or N may be infinite if the population characterized by X can take on an infinite number of values. Let $p_1, p_2, \ldots p_N$ be the probabilities that $X_1, X_2, \ldots X_N$ occur. Then we may define the expected value of X_i as:

$$E(X_i) = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i} = \sum_{i=1}^{N} p_i X_i \quad \text{since} \quad \sum_{i=1}^{N} p_i = 1.$$

Example: Suppose X = number of dots we observe on the upturned face of an unbiased die. X can then take on 6 possible values 1, 2, 3, 4, 5, 6. Thus our population is $X_1 = 1$, $X_2 = 2$, $X_3 = 3$, $X_4 = 4$, $X_5 = 5$, $X_6 = 6$. Since our die is unbiased, each face will be equally likely to turn up and the relative frequency with which it appears will be 1/6. Thus each $p_i = 1/6$ and $\sum_{i=1}^{6} = 1/6 + 1/6 + \ldots + 1/6 = 1$.

Then $E(X_i) = \sum_{i=1}^{6} p_i X_i = 1/6 \times 1 + 1/6 \times 2 + \ldots + 1/6 \times 6 = 3.5$.

Note that Ex_i in this case $= 1/6 \sum_{i=1}^{6} X_i = \mu$, the population mean.

An example, where the p_i are not all equal, arises when we consider 2 dice thrown together and X is the sum of the dots appearing on each of the faces of the two dice. Here there are $6 \times 6 = 36$ possible occurrences, but since certain occurrences result in the same observation, it is seen that we have only eleven different occurrences, i.e. 11 possible sums 2, 3, 4, 5, 6, ..., 12. Of the 36 possible occurrences, one results in a 2, a 1-spot on each die. Two result in a 3, a 1 spot on one die, a two dpot on the other, and vice-versa, etc, i.e. we get the following distribution of probability for the 11 sums observable:

Sum observed (X_i):	2	3	4	5	6	7	8	9	10	11	12
Relative frequency of occurrence (p_i):	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Notice that the $\sum_{i=1}^{N} p_i = 1$.

In this case $E(X_i) = \sum_{i=2}^{12} p_i X_i = 1/36 \times 2 + \ldots 1/36 \times 12 = 7$.

If our population containes an infinite number of observable values, then

$$E(X_i) = \sum_{i=1}^{\infty} p_i X_i = \mu, \text{ the population mean.}$$

Actually the above definition only applies to those variables that take on discrete values. If we have a continuous variable, an analogous definition, in terms of the integral calculus, would be used.

The population mean, μ, is by definition $\sum_{i=1}^{N} X_i p_i = E(X_i)$.

The population variance σ^2 is defined as

$$\frac{\sum_{i=1}^{N} (X_i - \mu)^2 p_i}{\sum_{i=1}^{N} p_i} = E(X_i - \mu)^2$$

Also $\sigma^2 = E(X_i - \mu)^2 = E(X_i - EX_i)^2$ since $\mu = EX_i$.

Rules of Operation with Expected Values:

1. $E(X_1 + X_2 + \ldots + X_k) = EX_1 + EX_2 + \ldots + EX_k$
2. $E(cX_i) = cEX_i$ where c is a constant.
3. $Ec = c$ where c is a constant.
4. $E\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} EX_i$

The proof of these rules follow from the definition of E and are left as an exercise.

Suppose that each variate X_i can be represented as the mean of the population plus some random deviation:

$$X_i = \mu + \epsilon_i.$$

Here we are assuming that the variation in the X_i is due to the variation in the ϵ_i and that a linear relationship between X_i and ϵ_i holds.

$$\bar{x}. = 1/n \sum_{i=1}^{n} X_i = 1/n \sum_{i=1}^{n} (\mu+\epsilon_i) = \mu + \sum_{i=1}^{n} \epsilon_i$$

Now $E\mu = \mu$

and $E\epsilon_i = 0$ since by definition $EX_i = E(\mu+\epsilon_i) = \mu$.

The variance of X_i then becomes $\sigma^2 = E(X_i-\mu)^2 = E(\mu+\epsilon_i-\mu)^2 = E(\epsilon_i)^2$ $= E(\epsilon_i)^2-(E\epsilon_i)^2 = E(\epsilon_i-E(\epsilon_i))^2 = \sigma\epsilon^2$; i.e. the variation in the X_i's can be accounted for by the variation in the random errors ϵ_i.

Another idea that will now be introduced is that of the independence of two variables X_i and X_j where $i \neq j$. X_i and X_j are said to be independent if the value that we observe for X_i is in no way influenced by the value observed for X_j, and vice-versa. I.e., if the probability that we jointly observe X_i and X_j is p_{ij} then this probability would be equal to $p_i p_j$ where p_i is the probability that X_i is observed and p_j is the probability that X_j is observed. Now the expected value of the joint observation $X_i X_j$ is given as

$$E(X_i X_j) = \sum\sum X_i X_j (p_{ij}).$$

If X_i and X_j are independent, then

$$\sum\sum X_i X_j p_{ij} = \sum\sum X_i X_j p_i p_j = (\sum X_i p_i)(\sum X_j p_j) = (EX_i)(EX_j)$$

Similarly 2 random errors ϵ_i and ϵ_j are said to be independent of one another if $E(\epsilon_i \epsilon_j) = (E\epsilon_i)(E\epsilon_j) = 0$. Since X_i depends upon ϵ_i and X_j depends upon ϵ_j we can state that X_i and X_j are independent of each other if and only if $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$.

Problem: If $X_i = \mu + \epsilon_i$ and $X_j = \mu + \epsilon_j$ show that $E(X_i X_j) = (EX_i)(EX_j)$ when $E(\epsilon_i \epsilon_j) = 0$.

We will now show that if $X_1, X_2, \ldots, X_n$ are n independent observations from a population with mean μ and variance σ^2, then the variance of the mean $\sigma^2_{\bar{x}} = \sigma^2/n$.

Assume the model $X_i = \mu + \epsilon_i$. Then

$$\bar{x}. = 1/n \sum_{i=1}^{n} X_i = 1/n \sum_{i=1}^{n} \mu + 1/n \sum_{i=1}^{n} \epsilon_i = \mu + 1/n \sum_{i=1}^{n} \epsilon_i$$

$$\sigma^2_{\bar{x}} = E(\bar{X}_i - \mu)^2 = E(\mu + \sum_{i=1}^{n} \epsilon_i/n - \mu)^2 = E(\sum_{i=1}^{n} \epsilon_i/n)^2$$

$$= 1/n^2 \, E \quad (\epsilon_1+\epsilon_2+\ldots+\epsilon_n)^2$$

$$= 1/n^2 \, E \quad (\epsilon_1^2+\epsilon_2^2+\ldots+\epsilon_n^2+2\epsilon_1\epsilon_2+\ldots+2\epsilon_{n-1}\epsilon_n)$$

$$= 1/n^2(E\epsilon_1^2+E\epsilon_2^2+\ldots+E\epsilon_n^2+2E\epsilon_1\epsilon_2+\ldots+2E\epsilon_{n-1}\epsilon_n)$$

$$= 1/n^2(\sigma^2+\sigma^2+\ldots+\sigma^2+2(0)+\ldots+2(0))$$

$$= 1/n^2(n\sigma^2) = \sigma^2/n$$

To show that : $E(s^2) = \sigma^2$ where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$ and $X_1 \ldots X_n$ are n independent variables drawn from a population with mean μ and variance σ^2.

$$\sum_{i=1}^{n}(x_i - \bar{x}.)^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n = \sum_{i=1}^{n}(\mu+\epsilon_i)^2 - 1/n\left(n\mu + \sum_{i=1}^{n}\epsilon_i\right)^2$$

$$E\sum_{i=1}^{n}(x_i - \bar{x}.)^2 = \sum_{i=1}^{n} E(\mu+\epsilon_i)^2 - 1/nE\left(n\mu + \sum_{i=1}^{n}\epsilon_i\right)^2$$

$$= \sum_{i=1}^{n}(E\mu^2 + 2E\mu\epsilon_i + E\epsilon_i^2) - 1/n\left(En^2\mu^2 + 2En\mu\sum_{i=1}^{n}\epsilon_i + E\sum_{i=1}^{n}\epsilon_i^2\right)$$

$$= n\mu^2 + 0 + n\sigma^2 - n\mu^2 - 0 - \sigma^2$$

$$= n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

or $$\frac{1}{n-1} E\sum_{i=1}^{n}(x_i - \bar{x})^2 = E\,\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = E(s^2) = \sigma^2.$$

One-way classification with unequal numbers of observations in each class.

In this case, the number of observations in the columns is not necessarily equal, i.e. we have k columns and n_j observations in the j-th column. The observations set out in a two-way table would look as follows:

x_{11}	x_{12} ...	x_{1j} ...	x_{1k}
x_{21}	x_{22}	x_{2j}	x_{2k}
x_{31}	x_{32}	x_{3j}	x_{3k}
$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_j j}$	$x_{n_k k}$
sums $X_{.1}$	$X_{.2}$	$X_{.j}$	$X_{.k}$
means $\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.j}$	$\bar{x}_{.k}$

where $X_{.j} = \sum_{i=1}^{n_j} X_{ij}$

$$\bar{x}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

$$\bar{x}_{..} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}}{\sum_{j=1}^{k} n_j} = \frac{\sum_{j=1}^{k} X_{.j}}{\sum_{j=1}^{k} n_j} = \frac{\sum_{j=1}^{k} n_j \bar{x}_{.j}}{\sum_{j=1}^{k} n_j}$$

Note that the total number of observations is $\sum_{j=1}^{k} n_j$. The breakdown of the total sum of squares can be given as follows:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{..})^2 = \sum_{j=1}^{k} n_j (\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})^2$$

For computational purposes the total sum of squares and between column sum of squares reduce to:

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - x)^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - \frac{(X_{..})^2}{\sum_j n_j}$$

$$\sum_{j=1}^{n_j} n_j (\bar{x}_{.j} - \bar{x}_{..})^2 = \sum_{j=1}^{k} \frac{X_{.j}^2}{n_j} - \frac{(X_{..})^2}{\sum_j n_j}$$

Both the breakdown of the sums of squares and the derivation of the computational formulas are left as an exercise for the student.

Two-way classification. Breakdown of the sum of squares. Assuming that each observation is classified with respect to two variables, e.g. X_{ij} = observation on yield of wheat for i-th variety treated with j-th fertilizer type.

The deviations of the individual observations from the grand mean can be broken down as follows:

$$X_{ij} - \bar{X}.. = (\bar{X}_{i.} - \bar{X}..) + (\bar{X}_{.j} - \bar{X}..) + (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)$$

where $(\bar{X}_{i.} - \bar{X}..)$ = deviations of row means from the grand mean.

$(\bar{X}_{.j} - \bar{X}_{..})$ = deviations of column means from the grand mean.

$(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)$ = deviations of individuals from the grand mean after differences in row and column means have been removed, i.e. we correct each X_{ij} as follows:

$X_{ij} - (\bar{X}_{i.} - \bar{X}..) - (\bar{X}_{.j} - \bar{X}..)$. Then we take deviations of the corrected X_{ij} from the grand mean:

$$X_{ij} - (\bar{X}_{i.} - \bar{X}..) - (\bar{X}_{.j} - \bar{X}..) - \bar{X}.. = (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)$$

Squaring both sides of the equation and summing we get:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}..)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n}\left[(\bar{X}_{i.} - \bar{X}..) + (\bar{X}_{.j} - \bar{X}..) + (\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)\right]^2$$

The sum of squares on the left hand side is the total sum of squares. On the right hand side, grouping the first two terms together and squaring we get:

$$\sum_{j}\sum_{i}\Big[\left\{(\bar{X}_{i.} - \bar{X}..) + (\bar{X}_{.j} - \bar{X}..)\right\}^2 + 2\,(\bar{X}_{i.} - \bar{X}..)\,(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..) + (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)^2\Big]$$

Squaring the term in curly brackets and summing, we obtain after rearranging the order of the terms:

$$\sum_{j}\sum_{i}(\bar{X}_{i.} - \bar{X}..)^2 + \sum_{j}\sum_{i}(\bar{X}_{.j} - \bar{X}..)^2 + \sum\sum(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)^2$$

$$+ 2\sum_{j}\sum_{i}(\bar{X}_{i.} - \bar{X}..)(\bar{X}_{.j} - \bar{X}..)$$

$$+ 2\sum_{j}\sum_{i}(\bar{X}_{i.} - \bar{X}..)(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)$$

$$+ 2\sum_{j}\sum_{i}(\bar{X}_{.j} - \bar{X}..)(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}..)$$

The last three terms all disappear since:

$$2\sum_{ji}(\bar{X}_{.j}-\bar{\bar{X}}..)(\bar{X}_{i.}-\bar{\bar{X}}..) = 2\sum_{j}(\bar{X}_{.j}-\bar{\bar{X}}..)\sum_{i}(\bar{X}_{i.}-\bar{\bar{X}}..) = 0,$$

$$2\sum_{ji}(\bar{X}_{i.}-\bar{\bar{X}}..)(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{\bar{X}}..) = 2\sum_{j}(\bar{X}_{i.}-\bar{\bar{X}}..)\sum_{i}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{\bar{X}}..) = 0,$$

$$2\sum_{ji}(\bar{X}_{.j}-\bar{\bar{X}}..)(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{\bar{X}}..) = 2\sum_{j}(\bar{X}_{.j}-\bar{\bar{X}}..)\sum_{i}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{\bar{X}}..) = 0.$$

therefore we have:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\bar{\bar{X}}..)^2 = k\sum_{i=1}^{n}(\bar{X}_{i.}-\bar{\bar{X}}..)^2 + n\sum_{j=1}^{k}(\bar{X}_{.j}-\bar{\bar{X}}..)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{\bar{X}}..)^2$$

The associated degrees of freedom can be gotten as follows:

1) For the total sum of squares we calculate the variation of nk observations with the restriction that the sum of the deviations of all observations from the grand mean must equal zero, i.e.

$\sum_{ij}(X_{ij}-\bar{\bar{X}}..) = 0$ therefore nk-1 d.f.'s.

2. For the sum of squares between row means we calculate the variation in n row means from the grand mean with the restriction that the deviations of the row means from the grand mean equals zero i.e. $\sum_{i}(\bar{X}_{i.}-\bar{\bar{X}}..) = 0$. Therefore we have n - 1 d.f.'s.

3. Similarily for column means we have the restriction $\sum_{j}(\bar{X}_{.j}-\bar{\bar{X}}..) = 0$. Therefore k-1 d.f.'s.

4. For the residual sum of squares we subtract the variation due to row and column means and therefore the degrees of freedom will be

$nk - (n-1) - (k-1) - 1 = nk - n - k + 1 = (n-1)(k-1)$.

Note that: $nk - 1 = (n-1) + (k-1) + (n-1)(k-1)$.